Performance Analysis of MPI over InfiniBand on Yellowstone

Zhengyang Liu
Mentor: Dr. John Dennis

Collaborators:
Prof. Malathi Veeraraghavan (University of Virginia)
Prof. Robert D. Russell (University of New Hampshire)
Fabrice Mizero (SIParCS)
Patrick MacArthur (University of New Hampshire)

Oct 25, 2013

Big Picture

- Understanding the causes of poor performance of CESM on Yellowstone: a 5-step approach
  - Experimental execution and data collection
  - CESM trace analysis
  - IBMgtSim: routing study
  - Network simulation
  - Integrated simulation
Big Picture

- Understanding the causes of poor performance of CESM on Yellowstone: a 5-step approach
  - Experimental execution and data collection
  - CESM trace analysis
  - IBMgtSim: routing study
  - Network simulation
  - Integrated simulation

Yellowstone network

251 A-groups
14 B-groups
9 ORCAs

*Credit: Dr. John Dennis*
Communication Patterns

- Intra-socket (via shared cache/memory)
- Inter-socket (via shared memory over QuickPath Interconnect)
- Inter-node (via InfiniBand)

Latency Benchmark: mpi_pingpong

- Approximate one-way latency by measuring round-trip latency
- Results represent ideal latencies between nodes
Yellowstone Results

- 0-byte messages
- 256 KB messages

Latency vs. # of Hops

- Experiment:
  - mpi_pingpong on 1024 cores
  - 1,048,576 communication pairs*
  - 256 KB messages

<table>
<thead>
<tr>
<th># of Hops</th>
<th># of pairs</th>
<th>Min.</th>
<th>Avg.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>15,680</td>
<td>5.95</td>
<td>52.20</td>
<td>88.29</td>
</tr>
<tr>
<td>2</td>
<td>59,904</td>
<td>47.72</td>
<td>48.83</td>
<td>60.64</td>
</tr>
<tr>
<td>4</td>
<td>588,736</td>
<td>49.55</td>
<td>53.30</td>
<td>114.10</td>
</tr>
<tr>
<td>6</td>
<td>332,016</td>
<td>52.75</td>
<td>56.92</td>
<td>159.30</td>
</tr>
</tbody>
</table>

*routing data available for 996,336 pairs
Bandwidth Benchmark: mpi_bw

- Measures throughput between two MPI ranks
- 3 communication patterns:
  - Intra-socket
  - Inter-socket
  - Inter-node
- 2 communication protocols:
  - Eager protocol
  - Rendezvous protocol

Communication Protocols

- Rendezvous Protocol: buffer negotiation before sending
- Eager Protocol: send directly without confirming available buffer space
- InfiniBand: Eager protocol uses SEND/RECV verbs (two-sided communication);
  Rendezvous protocol uses WRITE/READ verbs (one-sided communication)
- Eager Limit: threshold below which Eager protocol is used
Jellystone Results

- Intra-node throughput decreases when msg. size > eager limit
- Inter-node throughput increases when msg. size > eager limit
- Inter-node communication faster than inter-socket communication: RDMA vs shared memory

IBMgtSim and OpenSM

- Problem
  - Special version of OpenSM needed to boot a simulated network
  - Compile time configuration, not supported by Mellanox’s version
  - Open source version of OpenSM does not work with Mellanox IBUtils (which contains IBMgtSim)
  - Open source IBUtils outdated
- Solution
  - Fix bugs in open source version of OpenSM
  - Patches OpenSM to support FDR link rate
  - Patches to support better simulation work flow (RunSimTest)
Summary

- Identified contention through mpi_pingpong benchmarks
- Studied effect of different communication patterns/protocols on throughput

Future Work

- Analyses of larger data sets
  - > 500 million data points
  - Analysis needs to be parallelized
- Study interaction between MPI and InfiniBand
  - Open-source MPI implementations
  - Network sniffing
Thank You

Zhengyang Liu
zl4ef@virginia.edu